DOCUMENT RESUME

ED 270 484 TM 860 367

AUTHOR Shale, Doug
TITLE Essay Reliability: Form and Meaning.
PUB DATE Apr 86
NOTE 42p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Variance; Correlation; Error of Measurement; *Essay Tests; *Generalizability Theory; *Interrater Reliability; Measurement Techniques; Scores; Scoring; *Testing Problems; Test Interpretation; *Test Reliability; Writing Evaluation

ABSTRACT

This study is an attempt at a cohesive characterization of the concept of essay reliability. As such, it takes as a basic premise that previous and current practices in reporting reliability estimates for essay tests have certain shortcomings. The study provides an analysis of these shortcomings--partly to encourage a fuller understanding of the concept of reliability as applied to essay testing, and partly to build the case that the framework of generalizability theory offers a much more satisfactory way of characterizing the concept. The study applies generalizability theory to existing research to illustrate that this approach is an improvement over the usual methods of estimating essay reliability. The paper also argues that the classical approach to reliability has lead to a preoccupation with inter-marker agreement which in turn has led to a formulation of the "reliability problem" that makes it not susceptible of solution. Conceptual grounds for tolerating inter-marker disagreement are advanced and the paper discusses conditions under which this may be so. However, the paper explains how generalizability theory remains an appropriate framework for estimating the reliability of essay scores whatever assumptions one chooses to make regarding inter-marker consistency. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED270484

ESSAY RELIABILITY: FORM AND MEANING

Doug Shale
The University of Calgary

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. Shale

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

A paper presented at the Annual Meeting of the
American Educational Research Association
April 16-20, 1986; San Francisco

TM 860 367

ESSAY RELIABILITY: FORM AND MEANING

DOUG SHALE
THE UNIVERSITY OF CALGARY

ABSTRACT

This study is an attempt at a cohesive characterization of the concept of essay reliability. As such, it takes as a basic premise that previous and current practices in reporting reliability estimates for essay tests have certain shortcomings. The study provides an analysis of these shortcomings--partly to encourage a fuller understanding of the concept of reliability as applied to essay testing, and partly to build the case that the framework of generalizability theory offers a much more satisfactory way of characterizing the concept. The study applies generalizability theory to existing research to illustrate that this approach is an improvement over the usual methods of estimating essay reliability. The paper also argues that the classical approach to reliability has lead to a preoccupation with inter-marker agreement which in turn has lead to a formulation of the "reliability problem" that makes it not susceptible of solution. Conceptual grounds for tolerating inter-marker disagreement are advanced and the paper discusses conditions under which this may be so. However, the paper explains how generalizability theory remains an appropriate framework for estimating the reliability of essay scores whatever assumptions one chooses to make regarding inter-marker consistency.

ESSAY RELIABILITY: FORM AND MEANING

The concept of reliability has been well worked out for objective tests and seems well understood. Although the term often appears to be used loosely when applied to objective testing, it is usually apparent from context whether the reliability reported for a test has been estimated by correlating sets of scores derived from two administrations of the same test, scores derived from equivalent forms, scores derived from split-halves, or by analyzing components of variance. In turn, it is then usually apparent what meaning should be attached to the term "reliability."

However, the application of the concept of reliability to essay testing seems to have been considerably less clear (see, for example, McClearly, 1979, and the reply to this article by Thompson, 1980). Partly, this may be attributable to imprecise usage of the term--which may, in turn, reflect an imprecise understanding of what it is that possesses (or does not possess) the attribute "reliability." For example, in the literature, the term "reliability" has been applied to scoring procedures; to classes of scoring procedures (as in the "reliability" of direct/indirect assessment); to criteria used in analytic marking; to the "content" of essays; to essay topic; to essay scores; to the writers; and to inter- and intra-marker agreement. Considerable ambiguity may also arise because the full sense of reliability, as understood within the context of objective testing, may not transfer well--nor perhaps even appropriately--to the world of essay testing. (For example, what would constitute equivalent forms reliability or split-halves reliability in the context of essay testing?) Problems of interpretation have not been helped by speaking in a deterministic sense of <u>the</u> reliability, implying by this usage that there is just one common understanding for the term. There simply is "no single, universal and absolute reliability coefficient"

(Stanley, 1971). Moreover, as Cox (1969) has observed, "the use of 'reliability' in this context is not quite the same as its common use. It is used...to refer to the differences that occur when we repeat what is meant to be exactly the same measurement." He goes on to suggest that: "Even with this use, however, the term is still vague and there may be some point in dropping it and only thinking in terms of particular aspects of reliability like homogeneity, stability or marker variability."

Undoubtedly, the major factor responsible for the complexity of the concept of reliability in the context of essay testing is the subjective scoring process. It has long been known that marks awarded to essays may vary considerably from marker to marker when multiple markers are used, or from occasion to occasion when the same marker is used (see, for example, Hartog, Rhodes, and Burt, 1936; Hopkins, 1921; Huddleston, 1954). This has naturally lead to a concern that if different raters cannot agree in their ratings of the same essay, then the measurement procedure will not yield reliable results. The essence of the concern is that a satisfactory level of reliability of a measure is a necessary but not sufficient condition for the measure to be valid. Rater disagreement (or inconsistency) is an important consideration insofar as it acts as a limiting condition on the reliabilities of essay measures--and unreliability in these measures, in turn, acts as a limiting condition on validity.

For this reason, measures of inter-marker and intra-marker agreement have often been calculated and cited as indices of reliability. However, the use of indicators of rater consistency as estimates of reliability is likely to be confusing and misleading. As Finlayson (1951) among others has observed, rater consistency does not measure reliability in the usual sense of the word. Moreover, a pre-occupation with marking consistency may lead to its being

considered an end in itself (Stalnaker, 1951). Even in those instances in which it is explicitly recognized that the dependability of a measure is attenuated by inconsistency in rater judgements, there typically has been no direct calculation of the effect that rater disagreement has on the dependability of measures of essay scores--other than through a general eye-balling of whatever indice of marker agreement is available, and the application of some rule of thumb regarding what constitutes "too much" disagreement.

In passing it may be of some interest to note that the problem of rater variability is endemic to any situation in which human judgement is an integral component of a measure. It is a dominant theme in research predicated on observational measures (Rowley, 1976; Frick and Semmel, 1978). In the case of essay assessment, rater variability is not just a problem for English compositions. For example, Williams (1933) and Starch and Elliott (1913) demonstrated that wide variability can exist in marks awarded to solutions of mathematical problems. Head (1966) and Lucas (1971) have shown rater variability with respect to essay questions in Biology, and similarly Starch and Elliott (1913), and Modu (1972) with respect to Social Studies essays. In this paper, however, reference will typically be made to English composition because the bulk of research on essay reliability has been carried out in this area.

This study is an attempt to provide a comprehensive characterization of the concept of essay reliability. As such, it takes as a basic premise that previous and current practices in reporting reliability estimates for essay tests have shortcomings. The study will provide an analysis of these shortcomings--partly to encourage a fuller understanding of the concept of reliability as applied to essay testing, and partly to build the case that the framework of generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnam; 1972) offers a much more satisfactory way of characterizing the concept. In

the process of elaborating on generalizability theory, it will be particularly important to discuss in some depth the particular issues of marker agreement and essay topic.

The Concept of Reliability

The APA Standards for Educational and Psychological Tests and Manuals (1966; p.89) states: "Reliability refers to the accuracy (consistency and stability) of measurement by a test. Any direct measurement of such consistency obviously calls for a comparison between at least two measurements." Frick and Semmel (1978, p.158) have succintly re-stated this definition to: "Reliability has been defined classically as the consistency with which something is measured by maximally similar methods." In practice, the obtaining of at least two measurements has typically been accomplished by retesting an individual on the same test, by the use of parallel forms of a test, or through measures of internal consistency. It needs to be recognized, however, that the computation of a reliability coefficient depends not only on the test itself, but also on such factors as the characteristics of the individuals being tested and the conditions under which the test is administered--and, more generally, the sources of error that are taken into consideration by the particular reliability estimate being used (see Stanley, 1971, or the APA Standards for a comprehensive summary of this matter). To state the point in a slightly different manner, "... different methods of determining the reliability coefficient take account of different sources of error" (APA Standards, 1966; p.89). The APA Standards (1966; p.90) go on to state:

> "Hence, "reliability coefficient" is a generic term referring to various types of evidence; each type of evidence suggests a different meaning. It is essential that the method used to derive any reliability coefficient should be clearly described."

In the case of a single objectively scored test, it is apparent that the term "reliability" refers to a set of scores obtained on some particular

sample of examinees. Because such a test is used in such a constrained fashion, there is usually little confusion about what it is that the attribute "reliability" refers to. However, with essay tests, the presence of a rater (or raters) has clouded the issue somewhat. Consequently, in this paper the term "reliability," will refer to a measure (or, alternatively, a measurement) --a measure being a procedure for producing a score for each examinee. Characterizing reliability in this way has a number of advantages. In particular, it underscores the fact that: "A single instrument can produce scores which are reliable, and other scores which are unreliable. Even one measure may be reliable or unreliable, depending on the manner in which the instrument is used, the subjects observed..." (Rowley; 1976, p.53). Thus, it calls attention to the fact that it is essential to specify the conditions constituting the procedure whereby an instrument is used. One aspect of the procedure whereby measures are generated from essays would, of course, have to do with markers and marking conditions. It follows, then, that it would be extremely useful to have some structural theory to support this specification of conditions. The argument will be made that generalizability theory provides such a structure.

As indicated above, "...the formal definition of reliability has been phrased in terms of the correlation between parallel sets of measures..." (Stanley, 1971). However, there are a number of important assumptions inherent in this parallel measures concept that need to be mentioned because they bear on the meaning and sufficiency of actual calculations of reliability coefficients. Two (or more) tests are assumed to be parallel if they are equivalent in content, means, variance, and intercorrelations of items (Cronbach, Rajaratnam, Gleser; 1963). In practice, this is widely regarded as a very stringent definition of "parallel" that is seldom satisfied even when

traditional psychological tests are used. Moreover, as Stanley (1971) has observed: "... it seems obvious that the procedure for reliability determination which makes use of parallel forms will measure up to logical requirements. This is in fact true, provided satisfactory procedures for preparing parallel tests can be established." Whether or not satisfactory procedures can be established is arguable, and more will be said on this point later.

Calculating a reliability coefficient for a measure is a matter of determining, "... how much of the variation in the set of scores is due to certain systematic differences among the individuals in the group and how much to other sources of variation that are considered, for particular purposes, errors of measurement." (Stanley, 1971). In the classical formulation, an observed score is seen to be the sum of a "true score" and a purely random "error," with the error being regarded as a sample from a single undifferentiated distribution. In the classical context, a reliability coefficient is estimated as the ratio of the variance of true scores to the variance of observed scores--or, alternatively, it's interpreted "... as an estimate of the squared correlation of observed score with true score" (Cronbach, et al, 1972). The essential issue to be faced in determining the reliability of a measure relative to a particular group of individuals is defining what should be thought of as true variance among individuals and what should be thought of as error variance, given the investigator's purpose.

The hypothetical true score for an individual is classically defined as the average score that would result if a very large number of measurements were obtained on similar tests under similar conditions for the same individual. However, as Cronbach, et al (1972) point out: "(A) difficulty with the term "true score" is that the statistical concept of a limiting value approached through extensive observation is readily confused with some

underlying in-the-eye-of-God reality." (Sutcliffe, 1965, has referred to this "reality" as a "Platonic" concept of true score). In effect, a true score "... is the score resulting from all systematic factors one chooses to aggregate, including any systematic biasing factors that may produce systematic incorrectness in the scores" (Stanley, 1971). Stanley goes on to say that, "The heart of any treatment of reliability involves recognition that the true variance is wanted variance and that what is wanted will depend on the interpretation proposed by the investigator." Specifying what would constitute "similar tests" and "similar conditions" given the systematic factors one chooses to aggregate is problematic. In classical theory, the concept of parallel forms and parallel measurements provides a framework for addressing this issue. However, when a measurement procedure is based on the essay, the matter becomes even more complicated. Gosling (1966) approached this dilemma from the classical measurement point of view by defining a candidate's true score: "his mean score on an infinite number of tests;" and a script's true score: "the mean of the marks (awarded a script) by all qualified judges." He concludes: "ideally, then, to arrive at a perfect assessment of the ability we wish to measure, we should administer an infinite number of tests, each marked by all qualified judges" (p.8). However, such an approach begs two critical conceptual questions. In the first instance, there simply is no satisfactory theory to support the construction of parallel essay forms. Secondly, how should one regard the effect due to the introduction of markers as a condition of measurement? Are they "parallel instruments" in the same sense that different forms of a test can be parallel? Both of these points will be elaborated on in later sections.

On the other side, it has long been recognized that there are many ways to define "error." In general, though, error variance is that variation in a

particular set of measurements that will not be reproduced on another occasion. This error variance is a consequence of unreliability in a measurement procedure, and the magnitude of this variance relative to total variance is a measure of the unreliability. However, as Stanley (1971) points out, "Not every type of error, not every discrepancy from the "true" value that would register for the specimen in question, qualifies as a part of the error variance." What appears as error variance in a set of measures depends on how the measure is defined and on how the measurement is carried out. Thorndike (1947) has offered the following general classification of types of variance that can contribute to "error:"

1. Lasting and general.
2. Lasting but specific.
3. Temporary but general.
4. Temporary and specific.
5. Other.

The variance that is "wanted" with respect to individual differences is usually the lasting-general variance. The "other" category of errors consists of chance errors of measurement, "... that are unrelated to the future performance of the individual to which one wants to make inference" (Stanley, 1971). This is random error--the type of error that typically is of most concern. This "error" is usually regarded as being uncorrelated with an individual's true score or with the error of measurement for that person on another form of the test.

In the case of essays, Coffman (1972, p.7) points out:

> "The sources of error in essay examinations are complex. Some error arises because the questions in an examination are only a sample of all the possible questions that might be asked. Some error is the result of differences between raters. Some is due to the variability in the judgements of a rater from one time to another. Both interindividual and intraindividual variability can be further

> broken down into at least three components. The extent to which any of these various sources of error are present depends on how the essay questions are prepared, on how the responses are rated, and on how the scores are used."

The extent to which each of such sources of error will be of concern will depend upon the aims of the investigator (and, concomitantly, the design of the study).

The Correlational Approach

As mentioned previously, the traditional approach to assessing reliability is to calculate the correlation between two sets of scores (or measures) that, theoretically, are produced by "maximally similar methods." Given this coefficient of reliability and the standard deviation of the distribution of observed scores, an estimate can be derived for the standard deviation of the errors of measurement (i.e. the standard error of measurement). Again, under this approach, it is postulated that measures are strictly "parallel" such that test forms have equal means and variances and there is no interaction of subject with test form --as well as the instruments being equivalent in content. Moreover, variance is considered to arise as a result of "true" differences among subjects combined with random variation among observations (i.e. error). While this approach is reasonable for carefully equated parallel forms of tests, it is not really appropriate for measures that depart from these assumptions or that are qualitatively different. For example, in the case of essay testing, markers may differ in the central tendency of the scores they assign (thus giving rise to a main effect for markers), in the extent to which they differ in distributing scores across the grading scale being used, and in the attributes they attend to (thus giving rise to a subject-marker interaction). As Cronbach (1970) points out, the assumption of parallel measurement procedures, "... is not likely true for work samples, observations, or ratings from different acquaintances."

Breland (1983), in a comprehensive survey of the literature, has also made the point about there being a multitude of factors to be considered:

> "Reliability estimates (of direct assessments of writing skill) are influenced by the population studied, the number of cases examined, task type, number of tasks, number of readers, time allowed, scoring method used, and scoring range."

All of which to say there is an abundance of evidence that there are many sources of systematic variation possible in the case of essay measures. However, the classical approach to reliability cannot properly reflect these influences. As Coffman (1972) has stated "... a product-moment coefficient of correlation between two sets of ratings does not adequately assess all of these sources of errors in ratings. It takes into account only the fluctuations in relative standing." The classical approach deals with error variance as an amorphous single source of variance, typically leaving two or more sources entangled (Cronbach, 1970). Moreover, as mentioned earlier, where a study results in a correlation between two sets of measures, the value of the study is much enhanced if we can ascertain how much the observed correlation is attenuated by errors of measurement. In the case of essay measures, the usual major concern is how much effect does inconsistency in marking have on the dependability of the obtained measures of writing proficiency. Classical theory cannot address this question in any direct way. In addition, the standard correlational approach does not generalize well when more than two sets of measures are being considered.

Under the classical approach, any two sets of measures may be correlated with the result being expressed as a "reliability" coefficient. An additional shortcoming of this approach is that there is no easy way to differentiate the interpretation that ought to accompany each coefficient because it is not usually apparent what factors have contributed to the error variance. Consequently, these reliability coefficients may appear nominally similar, perhaps

even interchangeable. To state this in a converse form, coefficients that seem nominally the same, or similar, may in fact be extremely diverse with respect to the information on which they are based--and sometimes such coefficients will be mutually contradictory.

Finally, as Hopkins (1984) has observed, classical measurement theory has "... ignored latent random effects in the relevant universe of inference." Consequently, "... classical test theory ordinarily underestimates the degree of measurement error in the appropriate universe of generalization; that is, the inferences are not statistically congruent with those addressed by the reliability coefficient because undefined random sources of variation in the system are not acknowledged."

The Analysis of Variance Approach

As mentioned earlier, under the classical measurement approach the observed score is regarded as the sum of the true score and an error component, and reliability is estimated as the ratio of the variance of true scores to the variance of observed scores. It has long been recognized that analysis of variance (ANOVA) procedures could be applied to the estimation of components of variance (see, for example, Burt, 1936; Hoyt, 1941). An early example of this approach is the two factor (persons by items) ANOVA framework (originally presented by Hoyt, 1941, and re-presented in Cardinet et al, 1976) in which it was assumed that a sample of items was randomly drawn from a population of items, and a sample of persons was randomly drawn from a population of persons. Under this random-effects, persons-by-items model, the variance component for persons is estimated as the mean of the covariances among the items sampled. It is then possible to estimate an intraclass coefficient of correlation "... from the variance of the means of persons (which) is simply the ratio of the mean covariances among items to the mean of

the item variances" (Stanley, 1971). Stanley (1971) goes on to point out that under certain assumptions the classical correlational approach and the ANOVA approach coincide: "If the items are equally variable in the population of persons, ρ intraclass reduces to the mean Pearson product-moment ρ (among items), revealing that this intraclass coefficient of correlation is closely related to the mean interclass correlation of the I items in the form." This has also been demonstrated by Cronbach, et al (1972).

Gulliksen (1950) applied this approach to a study in which two essays per person were each graded twice. Using correlations derived from the two essay conditions and two different markers, Gulliksen formulated an estimate of the reliability of an essay test corrected for attenuation due to inconsistency in marking--a quantity that he called "the content reliability of the essay test," defining content reliability as "... the correlation between parallel forms divided by the geometric mean of the reader reliabilities of the two forms." Again, though, it should be pointed out that the classical assumptions on which the formulation was based are quite restrictive and there is no differentiation of the variability due to error sources.

In general, though, the introduction of Fisher's (1925) analysis of variance:

> "... revolutionized statistical thinking with the concept of the factorial experiment in which the conditions of observation are classified in several respects. Investigators who adopt Fisher's line of thought must abandon the concept of undifferentiated error. The error formerly seen as amorphous is now attributed to multiple sources, and a suitable experiment can estimate how much variation arises from each controllable source." (Cronbach, Gleser, Nanda, and Rajaratnam, 1972, p.1).

In spite of the obvious power of the analysis of variance framework, it has not seen extensive use in research based on essay measures. An early application of analysis of variance in essay testing is to be found in a study reported by Cast (1939). She used the procedure in a fully crossed, two

factor ANOVA design to examine: (i) a variance due to differences in merit between candidates, i.e. to variation in the average mark allotted to each candidate; (ii) a variance due to difference in the standard adopted by the several examiners i.e. to variation in the average mark that each examiner tends to award; and (iii) a residual variance due to random errors. The object of the study was to estimate the size of the three variance components and to test for significance of the two main effects. However, she did not make the link between estimating components of variance and formulating estimates of coefficients of reliability. The general relationship between variance components and reliability estimates was subsequently pointed out by Hoyt (1941), Jackson and Ferguson (1941); Lindquist (1953), Burt (1955), and others.

Finlayson (1951), using an analysis of components of variance worked out by Pilliner (1952), also used a factorial ANOVA design to examine the relative effects of markers, essays and writers. However, in his study, he extended the use of the variance components to calculate a reliability estimate for "an essay examination marked by the same markers;" and a reliability estimate for "the reliability of markers." By using the appropriate components, he obtained reliability estimates for each of the essays used and for each of the markers.

Pilliner (1952) set out in extended form the detailed development on which Finlayson's (1951) work was based. In particular, Pilliner presents formulations of ratios of combined variance components that result in reliability estimates for when: (i) "the same N children write essays on the same n topics as before, and each essay is marked by the same M markers as before;" (ii) "if the conditions are as above, except that a different set of markers is concerned with the second examination;" (iii) "the essays differ on the two

occasions and the markers remain constant;" and (iv) "if both essays and markers differ on the two occasions." Pilliner also compared the correlational approach and the ANOVA approach and demonstrated the particular conditions under which they converge. Wiseman (1956) used Pilliner's (1952) methods in an inquiry into essay reliability and validity.

Stanley (1962), capitalizing on extensions of the ANOVA technique to mixed models (wherein various effects may be regarded as fixed or random), examined the situation of two comparable (i.e. equivalent) essay forms, written by I examinees (who were assumed to have been drawn randomly from a population of examinees), and marked by R readers (who analogously were presumed to have been drawn randomly from a population of readers). Stanley presents ratios of variance components that provide estimates of: (i) how highly the two forms agree; (ii) "the mean reliability coefficient when Essay A is graded by a given reader and Essay B by the same reader;" (iii) "the average agreement among readers of the same form;" (iv) "the correlation between one reader's grades on one form and another reader's grades on a comparable form." In passing, it may also be of some interest to note that Stanley presents an F ratio that tests whether the two essay forms are measuring exactly the same thing.

In sum, the ANOVA approach represents a major advance over the classical correlational approach in characterizing the reliability of measures because: (1) it allows us to represent multiple sources of variation systematically in an experimental design and to estimate the effect of each--rather than dealing with such variance as an undifferentiated, amorphous whole; (2) this in turn permits a more refined appreciation of those factors that affect our sense of true variance and error variance; that is, our understanding is deepened by tracing how a change in study design affects each component of observed score

variance; and (3) having specific components of variance permits the formulation of different estimates of reliability that are appropriate to particular research questions.

However, an approach based strictly on the ANOVA framework also has shortcomings. First of all, the traditional ANOVA approach emphasizes testing for statistical significance of experimental effects--although, as is apparent from the work cited above, this needn't necessarily be the case. Furthermore, the classical applications of ANOVA in agricultural studies permitted direct measurement of variables and an exhaustive enumeration of experimental conditions--conditions which rarely, if ever, prevail in social science research. As Hopkins (1984) points out: "The numbers associated with bushels, pounds, pigs/litter, and so forth differ fundamentally from cognitive and affective measures in ways that have important implications for statistical analysis and interpretation. Items on tests and inventories are only a sample of the universe of items to which an inference is intended, whereas there is no sampling in the agricultural measures per se." Hopkins (1984) goes on to point out that the development of the distribution theory for "fixed," "random," and mixed ANOVA models has done much to improve matters. However, the analysis of variance framework is "passive" with regard to how to generalize appropriately from a particular study. That is, it is left entirely up to the investigator to be aware <u>at the time of designing a study</u> as to whether or not a particular factor should be designated as fixed or random--and what the inferential implications are as a result of this specification. As a result, as Hopkins (1984) points out: "Inferences drawn in education research are frequently not congruent with the statistical analysis because an important source of error is hidden and ignored in the statistical model employed. That is, a factor is implicitly treated as a fixed effect, yet the results are

interpreted as if the factor had been employed as a random factor." Moreover, unless one attends closely to the details of the design of a study, it will not be apparent from the results of an ANOVA just how the factors involved ought to be regarded and--most importantly--how the results should be interpreted. For example, the variance components given in Pilliner's (1952) analysis of a three factor design (individuals by markers by essay form)--and, in particular, the estimates of coefficients of reliabilities based on these components--differ from those given in Stanley's (1962) analysis of the same three factors. An examination of the details of the experimental design indicates that Stanley considered an effect with respect to the order within which the essays were written, whereas Pilliner didn't. Moreover, Stanley assumed the individuals and the markers used were samples from a population (that is, were random effects), whereas Pilliner seemingly regarded them as fixed. The important point to be made is that although Stanley and Pilliner present estimates of reliability that are nominally the same, in each case one would conduct the corresponding study differently--and consequently one also ought to qualify in some appropriate way the nature of the inferences that can be properly drawn in each case.

The analysis of variance framework <u>per se</u> does not support the elucidation of such distinctions. So, although ANOVA has provided the means for a more refined analysis of components of variance the problem still remains of illuminating "... the subtle inconsistency between the statistical analysis and the related universe of inference..." (Hopkins; 1984). Generalizability theory is widely viewed as a comprehensive structural theory for dealing with the dilemmas arising from classical reliability theory and the application of ANOVA procedures to estimating reliabilities. As Brennan (1983) points out: "To an extent, generalizability theory can be viewed both as an extension of

classical test theory and as an application of certain analysis of variance procedures to measurement models involving multiple sources of error."

Generalizability Theory

Brennan (1983), in his comprehensive introduction to generalizability theory, indicates some of the precursive work in this field. However, the foundations of the theory were explicitly set out in early work by Cronbach and his associates (1963; 1965), culminating in a book by Cronbach, Gleser, Nanda and Rajaratnam (1972) entitled The Dependability of Behavioral Measurements. Anyone interested in developing a reasonably full understanding of the field would be well advised to attend to the work by Brennan and by Cronbach et al (among others), for as Brennan points out, "... the power of generalizability theory is purchased at the price of some conceptual and statistical complexities." There is a very substantial literature regarding generalizability theory, and the account that follows here is necessarily abbreviated.

A behavioral measure has value only to the extent that it gives us information about some larger context. An observed score should at least be representative of the collection of measurements that might have been made--and the obtained score is of interest only because it tells us something about the expected value of other measures taken under equivalent conditions. This is the essence of the traditional understanding of the concept of "reliability." If an investigator could, he would measure a subject exhaustively over equivalent conditions and take the average over all measurements. As mentioned earlier, the average obtained through exhaustive measurement over equivalent conditions is referred to in classical measurement theory as "the true score." It has already been noted that the definition of "equivalent" condition is particularly problematic. Again, the true score is defined as

the expected mean score obtained under a particular set of circumstances--however, observed score variance is computed on scores obtained under a set of conditions which may be quite different. As Cardinet, et al, (1976) observe: "Using another set of questions, or repeating the same measure on another occasion, as was done traditionally, inevitably introduces sources of systematic variation." They go on to point out: "When conditions of a measurement situation can be maintained equivalent, the variability between one result and the next is likely to be limited. When measurement conditions are allowed to vary in one or several respects, the results are likely to be modified by the intervention of the corresponding sources of variation."

Under the requirement for equivalency, an investigator would need to determine the range of conditions over which true score variance ought to be estimated and would need to do so by some procedure that would yield reproducible results. But, as Cronbach (1970) has noted, there is no clear basis on which to do this.

Generalizability theory was formulated to address directly the conceptual difficulties inherent in the classical measurement approach. In generalizability theory, the concept of "true score" with its connotations of absoluteness is replaced by the concept of a "universe score," which Cronbach et al (1972) characterize as follows:

> "The score on which the decision is to be based is only one of many scores that might serve the same purpose. The decision maker is almost never interested in the response given to the particular stimulus objects or questions, to the particular tester, at the particular moment of testing. Some, at least, of these conditions of measurement could be altered without making the score any less acceptable to the decision maker. That is to say, there is a universe of observations, any of which would have yielded a usable basis for the decision. The ideal datum on which to base the decision would be something like the person's mean score over all acceptable observations, which we shall call his "universe score." The investigator uses the observed score or some function of it as if it were the universe score. That is, he generalizes from sample to universe. The question of "reliability" thus resolves into a

> question of accuracy of generalization, or generalizability. The universe of interest to the decision maker is defined when he tells us what observations would be equally acceptable for his purpose (i.e., would "give him the same information"). He must describe the acceptable set of observations in terms of the allowable conditions of measurement. This gives an operational definition to the class of procedures to be considered."

Under this conceptualization of universe score, generalizability (or "reliability") can be regarded as the expected value of the correlation between the set of obtained scores and other sets of obtained scores that could be drawn from the universe of interest. If the obtained scores derived from a particular measurement agree closely with the universe score, the observation may then be regarded as "reliable" or "generalizable." Because such observations must necessarily also agree well with one another, they are considered to be "dependable" or "consistent" and they will produce little error variance.

The counterpart of the traditional "reliability coefficient" is the coefficient of generalizability which is defined by Cronbach et al (1972) as "the ratio of universe-score variance to the expected observed-score variance. The size of the coefficient will be determined by the experimental design that has been followed in the study and will depend on the population of persons considered. In the framework of generalizability theory, a set of measures will not necessarily be reliable or unreliable in the traditional absolute sense; "... one can simply generalize, to different degrees, from one observed score to the multiple means of the different sets of possible observations. It follows that there are as many generalizability coefficients as sets of observations" (Cardinet, et al; 1976).

Cronbach (1970) coined the term generalizability "... because that term immediately implies 'generalization to what'?" As he points out, generalizing over scores is not the same thing as generalizing over passages. So, it's

important to note that a person will usually have a different universe score for each universe, and that there will typically be a different degree of generalizability for each universe. Consequently, the investigator has the considerable responsibility for defining the universe of concern to him because the universe of generalization is necessarily determined by him. Typically, an investigator should choose a set of observations in such a way as to ensure that they are representative of this universe. To state it in a converse form, the conditions under which measures are obtained will constrain the universe to which they can be generalized.

An investigator conducting a generalizability study will obtain two or more measures per person and will seek to determine how well they agree. Certain conditions (both known and unknown) will vary from one measurement to another and their influence will be accounted for in the "error" variance. Some other condition may be held constant from one measurement to another. If the analysis treats its effect as part of the universe-score variance, this would be incorrect unless the universe definition calls for holding that particular condition constant. As Cronbach (1970) notes: "An experiment that holds too much constant over estimates the universe-score variance, over estimates the coefficient, and under estimates the standard error of measurement." For example, if measures are obtained by having two raters mark every essay, the agreement of scores evaluates only one source or error--that originating from markers. In this case the coefficient of generalizability tells us how well from one marking we can generalize to the score a universe of judges would assign to that same performance. It tells us how well we have sampled judgements, but it cannot tell us anything about how well we have sampled a person's writing. An intraclass correlation among raters may be calculated, but it will ignore differences in marker means. Therefore, this

coefficient will be of relevance in just those studies in which markers will rate all essays. If, on another hand, a study is designed such that markers differ from essay to essay, the relevant intraclass coefficient in this case would be the one that treats marker leniency or severity as a source of error.

By this time it will be apparent that the variance components approach to reliability can become very complex, very quickly--and that "reliability" studies can correspondingly become very complex very quickly. Dealing with such complexity can require substantial, large-scale studies. Large-scale endeavours will be out of the question for most researchers. Moreover, for most researchers--although they necessarily have to be concerned with the "reliability" of the measures they obtain--the primary purpose of a study will be altogether different. However, generalizability theory does permit an indirect resolution of this problem. To describe the resolution, it will first be necessary to acquire some new terminology. In generalizability theory, an observation sampled from the universe of observations of interest to the investigator is characterized by the conditions under which the observation is made. In Cronbach's terminology, the set of all possible conditions of a particular kind is called a facet. As used in generalizability theory, the term "facet" "... serves to emphasize the distinction between the unit of analysis that is being observed and the facets, which indicate the conditions under which the observations are made" (Kane and Brennan; 1977). Within their theory, Cronbach, et al, distinguish generalizability (G) studies from decision (D) studies. The primary purpose of a G study is to collect data from which to derive estimates of the components of variance for measurements obtained by a particular procedure (i.e. to examine the dependability of some measurement procedure). Generally, "... G studies are most useful when they employ complex designs and large sample sizes to provide stable estimates of

as many variance components as possible" (Kane and Brennan; 1977). On the other hand, the primary purpose of a D study is to provide data on which to make decisions or draw conclusions. The components of variance obtained from a G study can be used to estimate coefficients of generalizability for various D study designs. This means, then, that the results of a G study with respect to the dependability of particular measures may be applied in D studies that may be addressing some other issue. However, this does presume at some time or another an appropriately designed G study must be conducted. It must be borne in mind, though, that the G study can serve the decision maker, "... only if its universe of admissible conditions is identical to or includes the proposed universe of generalization" (Cronbach, et al; 1972). It should be acknowledged that in some instances the same data may serve both G and D studies, for as Cronbach, et al (1972) point out, "The distinction between G and D studies is no more than a recognition that certain studies are carried out during the development of a measuring procedure, and then the procedure is put to use in other studies."

Although the line of argument in this paper has tended to present the various developments of the concept of "reliability" in a confrontational way, it ought to be stated that Cronbach and his associates viewed generalizability theory as a general framework within which to integrate previous definitions of "reliability." They demonstrated that these various definitions were not mutually exclusive and that they were, in fact, specific aspects of a more general model incorporating multiple sources of variation.

In spite of the considerable length of time that generalizability theory has been available, there has been remarkably little use made of it in writing research--to which it is remarkably well suited because of the multitude of variables that must be contended with. This author has been able to find only

one instance in which the methodology of generalizability theory has been applied in this field of study (Steele; 1979). In Steele's study, variance components were estimated for respondents, for raters (3 of them) and for writing tasks (3 of them). In this case, both raters and writing tasks were considered to be random samples from a larger universe of possible raters and a larger universe of writing tasks. The generalizability coefficient calculated for the marker effect was .70. An important result of his study, which will be referred to later (notwithstanding the irony in Steele's own violation of the spirit of generalizability betrayed in the following quote) was that, "... reducing the number of writing samples significantly reduces reliability, while increasing the number of samples beyond three or the number of raters beyond two does not seem to greatly enhance reliability."

Although the ANOVA studies cited earlier were originally approached from a traditional measurement point of view, one could retrospectively regard these studies through an overlay of generalizability theory. While the designs of these studies didn't explicitly attend to the issues of universe scores and universes of generalization, applying the framework of generalizability theory would serve to highlight the different assumptions made by the respective investigators. It would also serve to account for similarities and differences in the results reported for these studies. It may be of passing interest to note that the equations developed in these studies for estimating variance components--and the variance ratios presented as coefficients of reliability--may be entirely appropriate for other studies with facets established on the same assumptions.

As mentioned above, the facets in a generalizability study, acting either alone or in combination with one another, define universes--and the relevancy of such universes is determined by how the investigator proposes to interpret

the measure. Consequently, the design of a generalizability study will be particular to a particular investigator's purposes--and the equations for estimating variance components and related coefficients of "reliability" could vary considerably from study to study. There is no one set of equations to be offered for all purposes. However, Cronbach, et al, (1972) have worked through a number of complex designs that could be relevant to studies employing measures derived with essays--as has Brennan (1983).

Choosing an appropriate experimental design will largely be a matter of deciding which facets to incorporate into a generalizability study--and in what interpretive context. Because we know so very well that rater inconsistency can significantly affect measures obtained with essays, all studies in this domain ought to specify markers as a facet. We also know that individual writing performance can vary considerably from writing sample to writing sample (begging for the moment the issue of the nature of writing samples) --so, consequently, the study design should also include a facet for writing samples. Other facets will need to be introduced depending on the particular objectives or particular studies, but the facets of markers and writing samples will remain ubiquitous.

The inclusion of these two conditions as facets in generalizability studies using measures derived from essays seems reasonably unequivocal. However, identifying appropriate universes for these two facets seems to be much more problematic. Since we must be aware of what these universes are before we can select an appropriate experimental design, it may serve us well to look at the matter in some detail.

Universes of Writing Samples

Seldom will an investigator only be interested in a single piece of writing--typically the intent is to generalize to writing samples of a

particular kind, and perhaps occasionally to writing samples in general. However, there are a variety of universes of writing samples to which one might generalize and it is important to distinguish among them.

As a first cut at the problem, let us first distinguish the variability which arises when a measure is obtained for a person's performance on what is intended to be essentially the same writing task. This, in turn, traditionally was a matter of either responding to the same writing task on two (or more) occasions, or responding to two (or more) writing tasks that can be considered to be, in some sense, equivalent. This universe of generalization is generally regarded with limited--or qualified interest. First of all, in the case of multiple presentations of the same writing task, while it is true that by fixing the condition of rating task we may enhance the "reproducibility" of the measures obtained from it, we do so at the expense of the extent to which we can generalize the results of a study. And this is basically a point made by the APA Standards (1966), "Aside from practical limitations, retesting (an individual with the same test) is not a theoretically desirable method of determing a reliability coefficient if, as usual, the items that constitute the test are only one of many sets (actual or hypothetical) that might equally well have been used to measure the particular ability or trait." In the case of "equivalent" writing tasks, as noted earlier, there is the considerable difficulty of establishing formal equivalence (in the sense, that using either a particular writing task or its equivalent would not perturb the measures obtained). How does one go about prescribing conditions of equivalency in writing assignments--particularly with respect to content? Or is it even possible? Under the classic conceptualization of reliability, such equivalency was of course a vital concern.

However, if one regards the selection of a particular writing sample from the viewpoint of generalizability theory, matters become much more manageable. All one need do is address the question of what observations would be equally acceptable for the investigator's purpose. In this framework, we would be prepared to accept a specification that a particular writing task (the sample) should be like "these" (the universe of writing tasks). We are not forced to establish equivalency--we only need establish the basis on which a sample may be taken from the universe. This could then reduce to specifying a universe defined in terms of topics, in terms of mode of discourse, or whatever. However, because any particular set of equally acceptable observations identified in this way is likely to be quite heterogeneous (either by its inherent nature, or because of the varying interaction between writer and task) --selecting a sample of just one observation is not likely to be highly representative of the universe. Therefore, one ought to incorporate more than one writing task in any study using essays. To state the matter simply--we know that individuals performances on writing vary in response to writing task. Therefore, to get a more dependable measure for an individual, you should obtain observations on an individual for two or more writing tasks derived from the universe of interest.

Following this approach has considerable benefits. It rationalizes many practices that have in the past been shown to enhance the dependability of measures obtained with essays. For example, it has been demonstrated in a variety of ways that "reliability" (granting that the meaning of the term was inconstant) improves when the numbers of samples is increased (see, for example, Steele; 1979). It also highlights what is well known in sampling theory, but not widely appreciated in research in writing: results derived from samples of observations that are too small are very likely to be

unstable. Finally, such an approach can do much to mitigate the kinds of issues raised by Charney (1984), among others, regarding the selection of writing topics. For example, Charney, in the context of reliability of ratings, has asked: "Specifically, should writing samples representing different aims of discourse be compared?" The "answer" according to generalizability theory is: if an investigator determines that essay topics representing different modes of discourse are equally acceptable for his/her purpose, then it won't matter if writing tasks vary in this way. However, if this is so, then the investigator must, per force, be interested in some overall, general indication writing performance. And if this is so, clearly it will not be sufficient for the investigator to obtain a single writing sample representing only one mode. The proper sample should reflect the diversity inherent in the universe of interest--and consequently ought to be comprised of writing samples of all modes. If the observations were obtained for a writing sample representing only a single type of mode, then one is really only justified in generalizing to the universe defined by that particular mode.

However one chooses to address such matters as these, it is important to realize that there are, in fact, universes of writing tasks that we are sampling from. The implication of this is that writing task must be regarded as a "random" effect in a study design. Choosing to treat it as a "fixed" effect would be incorrect. To do so would yield an over-estimate of the universe-score variance, an over-estimate of the coefficient of generalizability and an under-estimate of the standard error of measurement.

Although generalizability theory may appear indifferent to the quality of writing assignments, this isn't so. The theory is robust enough to encompass poorly designed writing assignments as well as ones which are well designed. However, weak writing assignments will typically exacerbate the problem of

non-systematic variation between and among respondents. Even though we may be able to isolate and estimate error of this kind, the precision of our measures will diminish as the error term increases. Designing effective writing assignments, per se, is a matter beyond the scope of this paper. However, White (1985) offers a particularly good treatment of this issue and interested readers are referred to that source. It is important to note, though, that methods of designing effective writing assignments will necessarily reflect to a large degree, the universe of generalization that is of interest.

Universes of Markers

Much of what has been written regarding "unreliability" in essay testing has really been concerned with marker inconsistency--and much of this concern has been directed at inter-marker disagreement, and to a lesser extent intra-marker disagreement. As noted earlier, inconsistency in marking will necessarily attenuate the measures obtained with essays. However, calculating indices of rater inconsistency only tells us about rater disagreement, and there is a very substantial inferential step between this and the concomitant effect produced on the dependability of observed scores--an inferential step that has not been much aided by the measurement methodologies applied to date.

In passing, it may be worth elaborating on why measures of rater consistency are not measures of "reliability." Conceptually, regarding rater consistency as a measure of "reliability" results in (at least) two absurd conclusions. The usual one cited is that while mark-remark agreement has been regarded as an indication of reliability in essay testing, we would reject it out of hand for objective testing where such an index would, by definition, always equal one. The other line of thought would have us consider what it might take to achieve perfect mark-remark agreement. For example, one could diminish the potential for marker disagreement by setting an essay question

that would cause a more standardized response by examinees; one could force homogeneity in markers by selecting empirically those who agreed highly with one another; or one could produce a rating scale designed to force agreement by successively reducing in the numbers of categories in the marking scale by amalgamating them. As the number of categories diminished, marker agreement could be expected to increase. Ultimately, of course, one arrives at the absurd conclusion whereby there is only one category but perfect marker agreement--and a rating system that provides no information. In addition, from the viewpoint of generalizability theory it is apparent that introducing too many constraints distorts our sense of the dependability of sets of measures while at the same time limiting the extent to which we may generalize our results. All of these arguments have been advanced previously by various investigators. However, judging by the current literature, the importance of the distinction is not yet widely appreciated.

A major consequence of fixating on marker agreement has been that efforts to improve the "reliability" of essay measures have been directed primarily to ways and means of <u>eliminating</u> inter-marker variability. It is this author's opinion that this emphasis has lead to further distortions in measurement procedures based on essays. Such practices as analytic scoring have largely been introduced because they were viewed as a means of securing rater agreement--not because they were viewed as philosophically congruent with the aims of such measurement (and, in some instances, despite the fact that analytic marking schemes were viewed as antithetical to such aims). But perhaps the most questionable of the practices, is what in some cases has been an almost blind commitment to seeking homogeneous groups of markers--and in some instances creating such groups by expelling markers consistently deviating from group norms. There is a <u>reductio ad absurdum</u> inherent in the point of

view that inter-marker variability ought to be eliminated. We know from previous research that there are likely to be systematic differences among raters in the marks they award. It follows then that "error" variation will be greater with a group of raters than if a single rater were used. Consequently, we can obtain more dependable (i.e. "reliable") scores using just one marker. Clearly, this is not satisfactory because we must be haunted by the knowledge that had we selected a different rater (who also could have been highly self-consistent) we could have obtained quite different results.

However, there is a larger argument to be offered that eliminating inter-marker variability (as opposed to reducing it) is conceptually unsound. Much of this argument has been offered elsewhere, and will only be summarized here. Some of the argument is based on generalizability theory. Frick and Semmel (1978) - albeit in the context of observer agreement--cite previous work advancing the thesis "... that perfect observer agreement during actual data collection may not be particularly desirable. Since teachers and pupils in the real world do not always exhibit behaviors that neatly fall into predefined observational system categories, observer disagreement on ambiguities reveals a more representative picture of that real world." While acknowledging that this point of view may appear inconsistent, Frick and Semmel go on to elaborate:

> "It is highly improbable that any observation system has such specifically defined and perfectly mutually exclusive categories that every behavioral event that occurs can be clearly assigned to one of its categories. In all likelihood, there will be some teacher-pupil behaviors which are ambiguous--i.e., they cannot be clearly classified by a single category in the system. If observers are brainwashed to the point that they consistently code the same ambiguous behaviors into a certain category, results could be biased. Alternatively, if an observer codes an ambiguous behavior into one category and another observer codes the same ambiguous behavior into a different category, the overall results may indicate a more realistic description of the behavior of that teacher and/or pupil. That is, in the latter case there will be some tallies in

both categories, rather than in only one category as in the former case."

This will be even more so the case for essay measures, for as Hirsch (1977) has pointed out, throughout the history of literary evaluation "...critics have disagreed for centuries in their holistic judgments of texts, and, since the time of Plato and Aristotle, the fundamental grounds for their disagreements have been known. The structure of the problem has remained the same, in all of its many guises, throughout the centuries." He concludes that the "... problem of holistic assessment has been studied by some of the greatest thinkers of history. They have not solved the problem because it is not susceptible of solution. For that reason, and for purposes of research, we must restrict ourselves to judgments where agreement can be reached in principle, that is, to intrinsic judgment." To the present author, Hirsch's conclusions seem both unnecessary and unfounded. In order to elaborate on this assertion, and to explain how generalizability theory renders the problem "susceptible of solution," it will be necessary to digress briefly to review the distinction that Hirsch draws between intrinsic and extrinsic judgments of writing. Incidentally, Hirsch's book--and in particular the last chapter on assessing writing--is highly and unabashedly recommended by this author to those for whom such matters are of interest.

Hirsch characterizes extrinsic evaluation as a platonic mode of judgment, "... because it is based on criteria that are extrinsic to the writer's intentions, and even includes judgments about the quality of those intentions." Intrinsic evaluation, on the other hand, "... is a mode that begins and ends in the telos or implicit intention of the kind of writing that is judged. The quality of the text is judged according to its success in fulfilling its won implicit intentions, and these are not, by and large, to be measured against different intentions." Building on this distinction, Hirsch

identifies three deductively generated categories of writing assessment: (1) the quality of intentions; (2) the quality of their presentation; and (3) correctness. The quality of intentions, Hirsch subsumes under extrinsic evaluation. The quality of the presentation of intentions, and correctness are subsumed under intrinsic evaluation. According to Hirsch, judgments in general consist of some indeterminate amalgam of extrinsic and intrinsic evaluation--but it is the nature of extrinsic evaluation that is at the heart of the age old dilemma of disagreement in judgments of texts. He feels that by fixing attention on the intrinsic quality of presentation, we have, "... at least a sporting chance of solving the assessment problem--if the problem can be solved." As an aside, it's worth noting that he also proposes that matters of correctness be considered entirely separately.

To this author, Hirsch's emphasis on judging the quality of intentions may lead to the kind of situation that White (1985) reports has been deplored by Sommers (1982):

> "These markings almost universally treat the student text as simultaneously a finished product with editing faults and an unfinished part of the writing and thinking process. It is as if our confusion about evaluation is somehow bound up with a confusion about the nature of the student text, an odd form of literature created for the sole purpose of being criticized. Sommers finds that writing teachers tend to say the same things about student writing even though the texts in front of them change, as do the writers." (White; p.95).

Hirsch, as do many others, views "the assessment problem" as disagreement among judgments of text--and quite naturally he sees the elimination of such disagreement as the solution to this problem. The point of the preceding exposition was to indicate that such a point of view may be ill-founded on philosophical bases alone. From a measurement perspective, there are grounds for another resolution to the conundrum.

It is difficult to say for certain, but it seems as though the principle of eliminating rater variability has its roots in the classical measurement approach to "reliability." The classical approach is ill-suited to accommodating multiple sources of variation, so multiple sources of variation were eliminated (in theory, at any rate) by invoking strong conditions of equivalency. By this line of thought, then, a necessary condition for equivalency would be (in theory) identical marker behavior. The present author can find no logical or philosophical basis (independent of the measurement context) to support such a proposition. Does it not make more sense to accept that markers naturally vary in their judgments of texts (for a whole variety of reasons that can't be gone into here)--and to settle on a measurement structural theory that allows us to accommodate this as a reality? As indicated previously, generalizability theory provides this structure and permits us to specify markers as a facet in a study and estimate the variation that is due to markers--and to remove its effect from our considerations of other factors that may be of more direct interest.

Because we know from a substantial body of research that markers vary among themselves with respect to judgments of text, we must regard markers as a "random" effect--that is, the markers used in a study should be regarded as a sample drawn from a universe of markers. If each marker rates each essay, then the variance components estimates presented by Stanley (1962) may be relevant. To the extent that it is not possible a use a design in which every marker grades each essay, variance components estimates will have to be derived to accommodate whatever nesting or incompleteness may be present.

However, being able to accommodate variance due to markers should not be regarded as being "good enough." Marker variability, as noted earlier in the paper, will be comprised of variance that is systematic and variance that is

"random error" in the sense that it will not be reproduced over occasions. In general, the amount of random "error" will adversely affect coefficients of generalizability. Consequently, the amount of random "error" should be minimized insofar as possible. Up to a point, past practices in marker training have served this purpose (and to some extent so has the principle of selecting only "experienced" markers). In Hirsch's terms, it would be desirable to minimize variation with respect to the quality of the presentation of intentions, and the correctness of this presentation--and there are many ways in which this might (and should) be done. "Calibrating markers" (White; 1985) through such procedures will be especially important if we must regard a sample of markers to be selected from a large and undifferentiated population. However, other universes are possible. One of the most intriguing possibilities, and one which could address the issue of extrinsic evaluation raised by Hirsch is that of the "interpretive community." As White (1985; p.97) has pointed out:

> "Fish defines an interpretive community as made up of those whose common agreement about how to read texts becomes an agreement about how they will in fact "write" for themselves those texts: "Interpretive communities are made up of those who share interpretive strategies not for reading (in the conventional sense) but for writing texts, for constituting their properties and assigning their intentions. In other words, these strategies exist prior to the act of reading and therefore determine the shape of what is read rather than, as is usually assumed, the other way around" (1980, p.171). As Fish develops this concept, it serves a number of purposes. "This, then, is the explanation both for the stability of interpretation among different readers (they belong to the same community) and for the regularity with which a single reader will employ different interpretive strategies and thus make different texts (he belongs to different communities)" (p.171).

How one actually goes about specifying an interpretive community and identifying its constituents is a difficult problem. However, the notion of interpretive communities feels right, philosophically, and subsequent work may yet reveal its value in characterizing universes of generalizability.

There are two residual issues pertaining to marker consistency that it may be of value to elaborate on--one of these has to do with intra-marker agreement, and the other has to do with factor analytic approaches to the "reliability" of measures of writing performance. In the case of intra-marker consistency, the predisposition to eliminating marker variability seems reasonable--as do the many procedures that have been advocated for so doing. Nonetheless, a study design ought to give some consideration to this source of variance, if only to provide reassurance that it's not significant. In certain studies, it may also be essential to consider whether intra-marker variation may not be due to some systematic influence (particularly if marking occasions are separated by fairly long periods of time).

There is another approach to estimating the "reliability" of essay ratings that is based on factor analysis procedures. In the simplest form, one would examine the extent to which each marker loaded on a general factor--"reliability" being reflected in large loadings (see, for example, LaForge, 1965). In a more refined application, raters are regarded as a test instrument. Rater equivalence (which is then considered in analogous fashion to test equivalence) could then be of four kinds: parallel (in which the ratings of two raters have equal error variances and equal true scores); tau-equivalent (same true scores with possibly different error variances); congeneric (raters assign true scores that are perfectly linearly related); and a model in which error variances are assumed to be stable but true score variances vary from rater to rater. (For a more complete account see van der Kamp and Mellenburgh, 1976; Block, 1985). It is worth emphasizing the point that: "An assumption common to these four models is that for a set of essays, the true scores of one rater will correlate perfectly with the true scores of another rater" (Blok, 1985). Conceptually, this assumption is identical to

maintaining that inter-marker disagreement can be eliminated. The argument was advanced earlier in the paper that such an assumption is not well founded on philosophical grounds alone. Blok's (1985) finding that "... the ratings of different raters did not represent the same true scores," is congruent with this conceptual position.

Summary

This study is an attempt at a cohesive characterization of the concept of essay "reliability." As such, it has taken as a basic premise that previous and current practices in using reliability estimates for essay tests have certain shortcomings. The study attempted an analysis of these shortcoming--partly to encourage a fuller understanding of the concept of reliability as applied to essay testing, and partly to build the case that the framework of generalizability theory offers a much more satisfactory way of characterizing the concept. To this extent, the study has relied on existing research and an existing theory, and it has simply matched the two to illustrate that the product is an improvement over the usual characterizations of reliability.

However, the study also advances the argument that there are conceptual grounds for tolerating inter-marker disagreement. The paper discusses conditions under which this may be so, and explains how generalizability theory remains an appropriate framework for estimating the reliability of essay scores whatever assumption one chooses to make regarding inter-marker consistency.

REFERENCES

Blok, H. "Estimating the reliability, validity, and invalidity of essay ratings." Journal of Educational Measurement; 22(1); 1985; 41-52.

Breland, H.M. The Direct Assessment of Writing Skill: A Measurement Review. College Board Report No. 83-6; College Entrance Examination Board; New York; 1983.

Brennan, Robert L. Elements of Generalizability Theory. American College Testing Program; Iowa City, Iowa; 1983.

Burt, C. "The analysis of examination marks." In P. Hartog and E.C. Rhodes (Eds.), The Marks of Examiners. The Macmillan Company, London; 1936.

Burt, C. "Test reliability estimated by analysis of variance." British Journal of Statistical Psychology, 8, 1955, 103-118.

Cardinet, Jean, Tourneur, Yvan and Allal, Linda "The symmetry of generalizability theory: applications to educational measurement." Journal of Educational Measurement; 13(2); 1976; 119-135.

Cast, B.M.D. "The efficiency of different methods of marking English compositions." Parts I and II. British Journal of Educational Psychology; 1939, p.257-269; 1940, p.49-60.

Charney, Davida "The validity of using holistic scoring to evaluate writing: a critical overview." Research in the Teaching of English; Vol. 18, No. 1, 1984; 65-81.

Cox, Roy "Reliability and validity of examinations." In J.A. Lawreys and D.G. Scanlon (Eds.), World Book of Education, 1969: Examinations. Evans, London; 1969.

Cronbach, L.J., Rajaratnam, M., and Gleser, G. "Theory of generalizability: A liberation of reliability theory." British Journal of Statistical Psychology, 16, 1963, 137-163.

Cronbach, L.J. Essentials of Psychological Testing (3rd Edition). Harper and Row; New York; 1970.

Cronbach, L.J., Gleser, G.C., Nanda, H. and Rajaratnam, N. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. Wiley; New York, 1972.

Ebel, R.L. Essentials of Educational Measurement. Prentice-Hall, Inc.; New Jersey; 1972.

Finlayson, Douglas S. "The reliability of the marking of essays." British Journal of Educational Psychology; 21, 1951; 126-134.

References (Continued)

Fisher, R.A. Statistical Methods for Research Workers. Oliver and Bond, London; 1925.

Frick, Ted and Semmel, Melvin I. "Observer agreement and reliabilities of classroom observational measures." Review of Educational Research; 48(1), 1978; 157-184.

Gosling, G. Marking Compositions. Australian Council for Educational Research; Victoria, Australia; 1966.

Gulliksen, H. Theory of Mental Tests. John Wiley and Sons; New York; 1950.

Hartog, P.; Rhodes, E.C.; (with Burt, C.) The Marks of Examiners, Macmillan, New York; 1936.

Head, J.J. "Multiple marking of an essay item in experimental O-level Nuffield biological examinations." Educational Review, 19, 1966, 65-71.

Hirsch, Jr., E.D. The Philosophy of Composition. The University of Chicago Press, Chicago; 1977.

Hopkins, Kenneth D. "Generalizability theory and experimental design: incongruity between analysis and inference." American Educational Research Journal; 21(3), 1984; 703-712.

Hopkins, T.L. The Marking System of the College Entrance Examination Board. Harvard Monographs in Education, Series 1, No. 2. The Graduate School of Education, Harvard University; Cambridge, Mass.; 1921.

Hoyt, C.J. "Test reliability estimated by analysis of variance." Psychometrika, 6, 1941, 153-160.

Huddleston, E. "Measurement of Writing Ability at the College-Entrance Level: Objective vs. Subjective Testing Techniques." Journal of Experimental Psychology, 22, 1954; 165-213.

Jackson, R.W.B. and Ferguson, G. Studies on the Reliability of Tests. University of Toronto, 1941.

Kane, Michael T. and Brennan, Robert L. "The generalizability of class means." Review of Educational Research, 47(1); 1977, 267-292.

LaForge, R. "Components of Reliability." Psychometrika, 30, 1965, 187-195.

Lindquist, E.F. Design and Analysis of Experiments in Psychology and Education, Houghton Mifflin, Boston; 1953.

Lucas, A.M. "Multiple marking of a Matriculation Biology essay question." British Journal of Educational Psychology; 41, 1971, 78-84.

References (Continued)

McClearly, William J. "A note on reliability and validity problems in composition research." Research in the Teaching of English; 13, 1979, 274-277.

Modu, C.C. "The effectiveness of an essay section in the American history and social studies test." Educational Testing Service Research Bulletin, RB-72-5, 1972.

Rowley, Glenn L. "The reliability of observational measures." American Educational Research Journal, 13(1); 1976; 51-59.

Stalnaker, John M. "The essay type of examination." In E.F. Lindquist (Ed.), Educational Measurement. American Council on Education; Washington; 1951; 495-530.

Stanley, Julian C. "Analysis-of-variance principles applied to the grading of essay tests." Journal of Experimental Education; 30(3); 1962; 279-283.

Stanley, J.C. "Reliability," In R.L. Thorndike (Ed.) Educational Measurement. American Council on Education; Washington, 1971.

Starch, D. and Elliott, E.D. "Reliability of grading work in mathematics." School Review, 21, 1913; 254-259.

Starch, D. and Elliott, E.C. "Reliability of grading work in history." School Review, 21, 1913, 676-681.

Sutcliffe, J.P. "A probability model for errors of classification. I. General conditions." Psychometrika; 30; 1965; 73-96.

Thorndike, R.L. "Reliability." In A. Anastasi Testing Problems in Perspective, American Council on Education; Washington; 1967.

Vernon, P.E. and Millican, G.D. "A further study of the reliability of English essays." The British Journal of Statistical Psychology; Vol. VII, Pt. 11; 1954; 65-74.

van der Kamp, Leo J. Th. and Mellenbergh, Gideon J. "Agreement between raters." Educational and Psychological Measurement; 36; 1976; 311-317.

White, Edward M. Teaching and Assessing Writing. Jossey-Bass Publishers; San Francisco; 1985.

Williams, G.P. The Northampton Study Composition Scale; London; 1933.

Wiseman, Stephen "Symposium: The use of essays in selection at 11+.III. - Reliability and validity." British Journal of Educational Psychology; 26(3), 1956; 172-179.